Target article authors: Jonathan Phillips (corresponding author), Wesley Buckwalter, Fiery Cushman, Ori Friedman, Alia Martin, John Turri, Laurie Santos, & Joshua Knobe

Abstract word count: 60

Main text count: 1143 (with approval from the editor)

References word count: 257

Entire text word count: 1578

Full Commentary Title: Beyond knowledge vs. belief: The contents of mental-state representations and their underlying computations

Short Commentary Title: Contents and computations

Full names: Mika Asaba, Aaron Chuey, Hyowon Gweon

Institution: Stanford University, Department of Psychology

Full institutional mailing address:
450 Jane Stanford Way, Bldg. 420
Stanford University
Stanford, CA 94305

Institutional phone number: (650) 498-7832

Email addresses: masaba@stanford.edu; chuey@stanford.edu; hyo@stanford.edu

Homepage URL: https://sll.stanford.edu

Abstract:

Moving beyond distinguishing knowledge and beliefs, we propose two lines of inquiry for the next generation of ToM research: (1) characterizing the contents of different mental-state representations and (2) formalizing the computations that generate such contents. Studying how children reason about what others think of the self provides an illuminating window into the richness and flexibility of human social cognition.

Main text:

We agree with Philips et al. that examining a greater variety of epistemic states will enrich our understanding of the origins and development of mentalizing capacities, and appreciate their distinction between knowledge and belief. Importantly however, categorizing these mental states and asking at what age children can represent them are just the first steps towards characterizing the richness and flexibility of our social-cognitive capacities. Looking back on over a decade of research attempting to identify the earliest signatures of belief-attribution (e.g., Onishi & Baillargeon, 2005; Surian, Caldi, & Sperber, 2007; but see also Dörrenberg, Rakoczy, Liszkowski, 2018 and Powell, Hobbs, Bardis, Carey, & Saxe, 2018), we caution against yet another arms race to determine which representation is “more basic” or “more critical” to social learning. Instead, as cognitive scientists, we find ourselves asking: *How* are these mental state representations cognitively distinct from one another, and what cognitive mechanisms support these representations?

Imagine your roommate is watching as you try to open a jar that just won’t budge. You might say that your roommate *knows* you failed to open the jar, *believes* the jar is difficult to open, or even *thinks* that you are too weak to open it. While you would be comfortable using these words---know, believe, think---to describe your roommate’s mental states, their contents differ from what is typically studied in Theory of Mind literature; rather than reflecting verifiable external states of the world (e.g., the location of the jar), these mental states concern outcomes of intentional, goal-directed actions (e.g., failure to open a jar) and outputs of additional inferences based on observed action-outcome relationships, such as subjective evaluations about abstract qualities of objects (e.g., the jar is tricky to open) and agents, including oneself (e.g., your physical strength).

To understand how our mind flexibly generates, represents, and attributes these mental states, we need to move beyond traditional concepts and empirical paradigms that have dominated the past decades of research on Theory of Mind. From this perspective, young children’s reasoning about how others represent and infer abstract qualities of people (including the self) provides a particularly illuminating window into the richness and flexibility of human social cognition. Our recent work finds that children, by four years of age, are already capable of reasoning about what others think of them after observing their own failures or successes (Asaba & Gweon, 2018; Asaba & Gweon, under revision). Looking forward, we propose focusing our efforts on two related lines of inquiry: (1) characterizing the *contents* of mental states children (and non-human primates) can represent and (2) understanding and formalizing the *computations* (i.e., inferential processes) that give rise to such representations.

Contents of mental states If your roommate observed your failure to open the jar countless times every day, you may intuitively feel that your roommate “knows” you cannot open the jar. Similarly, we might use the word “know” to describe one’s various attitudes towards someone else (e.g., Sally knows that Ann is generous, funny, and competent”), especially when we suspect one has strong evidence about these qualities. However, these are inherently subjective evaluations that do not have objective, verifiable criteria for determining their truth value; they can only be expressed as the degree to which one “believes” X is true, rather than as a Boolean value (i.e., either true or false). Although people often describe these mental states using knowledge-laden language, these contents go beyond the scope of what Philips et al. would consider as knowledge.

Critically however, the content of these representations also differ from the content of beliefs that are typically studied in Theory of Mind literature. Rather than observable states of the physical world that are verifiable via perception (e.g., “Sally knows her toy is in the box” or “Sally thinks her toy is in the box”), these belief states concern abstract properties of agents that must be inferred from an agent’s behaviors or other social sources of information (e.g., others’ evaluative feedback or testimony, such as “Ann is very generous”). Beyond distinguishing knowledge vs. belief, we need more research on *why* children find some mental-state contents harder to attribute than others.

Underlying Computations Relatedly, the process by which we attribute mental states about internal qualities of agents may involve more complex computations than those concerning external states of the world. When an agent has direct perceptual access to a world-state, there is a one-to-one correspondence between what they see and what they represent. When the agent loses perceptual access while the world-state changes, the agent is rendered ignorant (i.e., Sally does not know where her toy is) or mistaken (i.e., Sally falsely believes that her toy is in the box). While understanding the relationship between an agent’s perception and their resulting epistemic state is already an impressive feat, representing others’ beliefs about internal qualities is even more so; these representations cannot be derived from perceptual access alone, and require further inferences based on an intuitive understanding of how observations of an agent’s goal-directed actions give rise to representations of the agent’s abstract qualities.

Much of the prior literature on Theory of Mind development has studied how children represent others’ ignorance or false beliefs that are decoupled from reality. However, these representations reflect only a fraction of the mental states we encounter in our everyday social interactions. When Sally observes Ann bake a delicious cake, get a “D” on a math exam, or donate $20 dollars to charity, what kind of mental states might children attribute to Sally? These representations could be about anyone, but they are especially powerful when they concern qualities of the self: Does Sally think I am good at baking? Terrible at math? Generous or stingy? Although we, as adults, naturally entertain these thoughts, more work is needed to understand how young children integrate their understanding of the physical and social world to attribute these nuanced mental states. Recent computational work has made major advances in formalizing the generative process by which an agent’s observation gives rise to beliefs about the external world (Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017; Jara-Ettinger, Schulz, & Tenenbaum, 2020); these approaches can provide important insights here as well.

Ultimately, the ability to represent others’ mental states is both a blessing and a curse. When directed at us, it inspires our motivation to learn and responsibility for our own actions; when gone awry, it dogs us with unnecessary worries about how others might evaluate us. Yet, for better or worse, we rely on these abilities to learn from others in a complex social world filled with competition, cooperation, and collaboration. The richness of human cultural knowledge comes from our ability to appreciate, evaluate, criticize, and communicate a host of abstract thoughts. By studying how the human mind supports these rich mental-state inferences, we can better understand how humans harness these capacities to learn from others and help others learn.

References

Asaba, M. & Gweon, H. (2018). Look, I can do it! Young children forego opportunities to teach others to demonstrate their own competence. In Kalish, C., Rau, M., Zhu, J., & Rogers, T. (Eds.), Proceedings of the 40th Annual Conference of the Cognitive Science Society (pp. 106-111). Austin, TX: Cognitive Science Society.

Asaba, M., & Gweon, H. (under revision). Young children rationally revise and maintain what others think of them. *Nature Communications.* Preprint DOI: 10.31234/osf.io/yxhv5

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1, 1–10. DOI: 10.1038/s41562-017-0064

Dörrenberg, S., Rakoczy, H., & Liszkowski, U. (2018). How (not) to measure infant Theory of Mind: Testing the replicability and validity of four non-verbal measures. *Cognitive Development,* 46, 12–30. DOI: 10.1016/j.cogdev.2018.01.001

Jara-Ettinger, J., Schulz, L. E., & Tenenbaum, J. B. (2020). The Naïve Utility Calculus as a unified, quantitative framework for action understanding. *Cognitive Psychology*, 123, 101334. DOI: 10.1016/j.cogpsych.2020.101334

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science, 308*(5719), 255-258. DOI: 10.1126/science.1107621

Powell, L. J., Hobbs, K., Bardis, A., Carey, S., & Saxe, R. (2018). Replications of implicit theory of mind tasks with varying representational demands. *Cognitive Development*, 46, 40–50. DOI: 10.1016/j.cogdev.2017.10.004

Surian, L., Caldi, S., & Sperber, D. (2007). Attribution of beliefs by 13-month-old infants. *Psychological science, 18*(7), 580-586. DOI: 10.1111/j.1467-9280.2007.01943.x